Title: PRESERVING PRIVACY WHILST MAINTAINING ROBUST EPIDEMIOLOGICAL PREDICTIONS

Author: Marleen Werkman Michael J. Tildesley Ellen Brooks-Pollock Matt J. Keeling

PII: S1755-4365(16)30034-2
DOI: http://dx.doi.org/doi:10.1016/j.epidem.2016.10.004
Reference: EPIDEM 223

To appear in:

Received date: 9-3-2016
Revised date: 10-10-2016
Accepted date: 12-10-2016

Please cite this article as: Werkman, Marleen, Tildesley, Michael J., Brooks-Pollock, Ellen, Keeling, Matt J., PRESERVING PRIVACY WHILST MAINTAINING ROBUST EPIDEMIOLOGICAL PREDICTIONS.Epidemics http://dx.doi.org/10.1016/j.epidem.2016.10.004

PRESERVING PRIVACY WHILST MAINTAINING ROBUST EPIDEMIOLOGICAL PREDICTIONS

Marleen Werkman[1,2], Michael J. Tildesley[1,3], Ellen Brooks-Pollock[4] Matt J. Keeling[1]

1 WIDER Centre, Mathematics Institute and School of Life Sciences, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK

2 Current address: Department of Infectious Disease Epidemiology, School of Public Health, Faculty of Medicine, St Marys Campus, Imperial College London, London, UK

3 Fogarty International Center, US National Institute of Health, Bethesda, MD 20892, USA

4 School of Social and Community Medicine, University of Bristol, Oakfield Grove, Clifton BS8 2BN, UK

Highlights

In particular, we have added the spatial results for Cumbria to the supplementary information. All changes that have been made in response to the reviews are detailed below. We feel that the addition of these analyses significantly strengthen the paper and make it suitable for publication in Epidemics.

ABSTRACT

Mathematical models are invaluable tools for quantifying potential epidemics and devising optimal control strategies in case of an outbreak. State-of-the-art models increasingly require detailed individual farm-based and sensitive data, which may not be available due to either lack of capacity for data collection or privacy concerns. However, in many situations, aggregated data are available for use. In this study, we systematically investigate the accuracy of predictions made by mathematical models initialised with varying data aggregations, using the UK 2001 Foot-and-Mouth Disease Epidemic as a case study. We consider the scenario when the only data available are aggregated into spatial grid cells, and develop a metapopulation model where individual farms in a single subpopulation are assumed to behave uniformly and transmit randomly. We also adapt this standard metapopulation model to capture heterogeneity in farm size and composition, using farm census data. Our results show that homogeneous models based on aggregated data overestimate final epidemic size but can perform well for predicting spatial spread. Recognising heterogeneity in farm sizes improves predictions of the final epidemic size, identifying risk areas, determining the likelihood of epidemic take-off and identifying the optimal control strategy. In conclusion, in cases where individual farm-based data are not available, models can still generate meaningful predictions, although care must be taken in their interpretation and use.

INTRODUCTION

Mathematical models form an integral part of epidemic preparedness planning and real-time forecasting [1-5]. State-of-the-art individual farm-based models involve detailed data: these data are often not available or, if available, are often not in the public domain owing to privacy concerns. However, sharing data will hugely benefit developing, optimising and training disease simulation models [6]. In some countries, only spatially aggregated data are available. However, the full heterogeneity of individual farms may not be captured with these data [7], and in cases where limited data are available, simpler models may be a necessity [8]. Striking the optimal balance between detail and utility is an open and problem specific question.

Individual farm-based models have been utilised in the past to aid in the understanding of epidemiological processes of Foot-and-Mouth Disease (FMD) and testing potential control strategies such as (ring) vaccination, culling of livestock and quarantine of infected premises, most notably during and in the aftermath of the UK 2001 epidemic [9-10]. These models typically rely on the availability of detailed spatial information regarding the size and location of all livestock farms [1]. Whilst these data are available for the UK, this is not the case for many countries around the world. For example in the USA, farm location data are aggregated at the county level to prevent privacy difficulties [11] and in Australia precise farm locations are not known for all states [12]. In many other countries around the world precise farm locations are not known at all.

In situations where detailed demographic data are not available but aggregated data are available, it may be possible to adopt a metapopulation approach when developing a mathematical model. Metapopulation models are often used in ecology, theoretical biology and epidemiology [13-18]. In metapopulation models, epidemiological units, such as farms, are spatially aggregated into patches or subpopulations. Within a patch, farms are assumed to be well mixed (in the sense that transmission occurs randomly between all pairs of farms, in a density-dependent manner) and behave uniformly. Transmission within and between patches must capture the physical processes and can occur via different routes and over different spatial scales, such as local processes (i.e. aerosol spread, direct contact of animals,

contaminated vehicles or farm equipment) or by long distance contact such as live animal movements [1,2,19,20].

Given that within a patch, farms are assumed to be well mixed and behave uniformly, a metapopulation model will not capture the impact of local spatial clustering of farms, heterogeneity of farm size nor species composition. Previous studies have shown that these characteristics may often play an important role in epidemic dynamics [1,21,22]. In this study, we therefore investigate whether, and under what circumstances, a metapopulation model is a good alternative to an individual farm-based (IFB) simulation model. Our goal is to determine whether a novel metapopulation model gives comparable predictions to the IFB model when considering key epidemiological quantities such as spatial spread, epidemic size and distribution of epidemic size. The results presented here will ultimately have implications for human and veterinary health settings where precise locations of farm are unknown.

2. MATERIALS AND METHODS

2.1 DATA AND MODEL

Information on farm locations, sizes and species compositions was obtained from the 2010 agricultural census provided by the Department of Environment, Food and Rural Affairs. Early versions of the model used in this paper assumed one single set of parameters for the whole UK [1]. However, a more accurate fit to the 2001 UK outbreak can be achieved by fitting individual parameter sets to five distinct regions of the UK – Cumbria, Devon, the rest of England (excluding Cumbria and Devon), Scotland and Wales [23]. This allows for the model to capture region-specific farming practices and control implementation. In particular, lower transmissibility values are found in Wales [23], possibly owing to the increase in road distances between farms (particularly in the hilly sheep farming regions). Previous work indicates that this regionalised model provides a more accurate fit to the 2001 outbreak than a model with a single set of parameters for the entire country [23]. In this paper, we consider outbreaks in Cumbria, Devon and Aberdeenshire and therefore utilise model parameters for these three counties (where Aberdeenshire parameters are fixed to those of Scotland).

2.2 WITHIN FARM DYNAMICS

Farms are classified as susceptible, exposed, infectious, reported or culled (SEIRC). The latent period (time for an exposed farm to become infectious) is set to five days, consistent with estimates from 2001 [1]. After this period, farms remain infectious for four days before being reported. There is then a two-day delay (during this period the farm remains infectious) from reporting to culling in line with previous work [1,9]. We adopt a Markovian approach for the transition between classes; these are modelled as a constant rate, leading to exponential distributed periods. In this paper, we assume that the virus spreads rapidly when introduced in a naïve farm, such that within-farm dynamics can be excluded from the model and all animals on a farm are assumed to belong to the same disease status (i.e. all animals on a farm are either susceptible or exposed etc.).

2.3 BETWEEN FARM DYNAMICS

2.3.1. INDIVIDUAL FARM-BASED (IFB) MODEL

In the 2001 IFB model, local spread, incorporating multiple routes of transmission (trucks, airborne transmission etc.) is modelled via the use of a distance dependent transmission kernel [1,8]. The local transmission kernel exhibits power-law like behaviour, such that that farms (j) that are in the closest proximity of an infected farm i experience the largest risk of transmission:

$$K\left(d_{ij}\right) = \frac{z}{1+\left(\frac{d_{ij}}{q}\right)^{s}}$$

Parameters s, q and z define the shape and scale of the kernel. These parameters are estimated from the 2001 local transmission kernel [1,11,24] and are set to 3, 1 and 0.12 respectively. The variable d_{ij} defines the Euclidean distance between any two farms i and j. In this model, the risk of infection is determined by the number of cattle and sheep on infected and susceptible farms and the Euclidean distance between them. The susceptibility (a) and infectivity (b) are species-specific and scale non-linearly with farm

size using parameters p and q. The stochastic rate of transmission from farm i to j is therefore given by:

$$\text{rate}_{ij} = \left(a_c N_{c,j}^{p_c} + a_s N_{s,j}^{p_s}\right) \left(b_c N_{c,i}^{q_c} + b_s N_{s,i}^{q_s}\right) K\left(d_{ij}\right)$$

where N represents the number of animals on a farm for cattle (c) or sheep (s).

Only cattle and sheep farms are included in this study; other species such as pigs are susceptible as well, but did not appear to play an important role during the 2001 and 2007 FMD outbreaks [20,25].

2.3.2. HOMEGENOUS METAPOPULATION MODEL

For the metapopulation model, the UK is divided into grids (two-dimensional squared cells) with an equal width and height. Farms are then allocated to the grid cells based on their Easting and Northing coordinates of the farmhouse taken from the cattle tracing system [5,19]. To investigate the effect of the resolution of the grid cells on model outcomes, we vary the scale of the grid cells from 200 metres (in order for most grid cells to contain only a single farm) to 10 km (with increments of 200 metres up to 1 km and 1 km increments from 1 km to 10 km grid cell sizes). The upper limit was chosen as this fits with many spatial scales used in control such as surveillance zones. As with most metapopulation models, we assume that all farms within a grid cell are well mixed.

In order to estimate the mean transmission rate within and between any two grid cells, we calculate the distance between two randomly located farms in each grid cell by integrating over all possible locations of farms of the two grid cells k and l:

$$\text{mean}\left(kernel_{kl}\right) = \frac{1}{|A_k||A_l|} \int_k \int_l K\left(\left\| x^l - y^k \right\|\right) dx\, dy$$

where A_k is the area of grid cell k and A_l is the area of grid cell l, and x and y refer to the point locations in grid cells l and k.

The susceptibility of grid cell l (Sus) and transmissibility of grid cell k ($Tran$) are species-specific and scale non-linearly with farm size using parameters p and q. The

transmissibility and susceptibility of a grid cell is the mean susceptibility (Sus) and transmissibility ($Tran$) of all farms (F) in grid cell k and l:

$$\text{mean}\left(Sus_l\right) = \frac{1}{F_l}\sum_{l=1}^{F_l}\left(a_c N_{c,l}^{p_c} + a_s N_{s,l}^{p_s}\right)$$

$$\text{mean}\left(Tran_k\right) = \frac{1}{F_k}\left(b_c N_{c,k}^{q_c} + b_s N_{s,k}^{q_s}\right)$$

The rate of transmission within grid cells and from grid cell k to grid cell l is therefore given by: $\text{rate}_{kl} = S_l \times \left(I_k + R_k\right) \times \text{mean}\left(kernel_{kl}\right) \times \text{mean}\left(Sus_l \times Tran_k\right)$

where S_l represents the number of susceptible farms in grid cell l, I_k and R_k represent the number of infected and reported in grid cell k respectively.

2.3.3 HETEROGENOUS METAPOPULATION MODEL

In order to investigate the transition between the homogeneous metapopulation model and the IFB model, we developed a metapopulation model with additional heterogeneous structure. Farm composition and size have been shown to play a major role in the course of an epidemic in other studies [1,19]. In order to investigate the effect of farm size and species composition upon epidemic spread, we adapt our homogenous metapopulation model described above to incorporate heterogeneity of farm size by aggregating the population within a grid cell into large and small farms. The threshold of small and large populations is chosen by calculating the median number of animals of each species on all sheep and cattle farms in the UK. Farms are therefore divided into four groups: (1) large cattle (>12.58 cows) and large sheep (>30 sheep) population, (2) large cattle population and small sheep population, (3) small cattle population and large sheep population and (4) small cattle and small sheep population. In this case we define the susceptibility and transmissibility of a grid cell as:

$$\text{mean}\left(Sus_n^l\right) = \frac{1}{F_n^l}\sum_{l_n=1}^{F_n^l}\left(a_c N_{c,l_n}^{p_c} + a_s N_{s,l_n}^{p_s}\right)$$

$$\text{mean}\left(Tran_m^k\right) = \frac{1}{F_m^k}\sum_{k=1}^{F_m^k}\left(b_c N_{c,k_m}^{q_c} + b_s N_{s,k_m}^{q_s}\right)$$

m and n represent the four groups of farm types for the infected and susceptible grid cell respectively. The rate at which farms in grid cell l becomes infected by grid cell k in this model is defined as:

$$\text{rate}_{mn} = \left(S_n^l \times Sus_n^l\right) \times \sum_m \left[\sum_k \left(I_m^k + R_m^k\right) \times Tran_m^k\right] \times \text{mean}\left(kernel_{kl}\right)$$

2.4 EFFECTS OF SPATIAL CLUSTERING

Spatial clustering is known to have a substantial effect on disease dynamics. We therefore investigate whether metapopulation models are still capable of capturing epidemic dynamics for various degrees of clustering. We create a synthetic dataset including 2500 farms in a 50 by 50 km square; this results in a comparable farm density as in Devon. Farm size and composition are selected at random from the Devon farms and also the Devon parameters are used to determine the susceptibility and transmissibility. Two extremes are compared, one where the farms are located randomly in a 50 by 50 km grid cell and one clustered scenario.

To generate the clustered synthetic farm locations, we use the same method as described in [21], whereby an algorithm was developed to create theoretical spatial distributions of farms. In [21], the average number of farms in a circle radius r around any given farm is given by a sum of exponentials such that:

$$D(r) = S_{\text{inf}} + \left(S_0 - S_{\text{inf}}\right)\left(\sum_i A_i e^{-B_i r}\right)$$

S_0 defines the average local density around a farm, the parameter S_{inf} defines the long-distance asymptotic density, whilst $\sum_i A_i = 1$ and the parameters B_i describe the different

length-scales of clustering that are observed. For the purposes of this analysis, we define the density distribution of farms in terms of a single exponential decay function such that:

$$D(r) = S_{inf} + (S_0 - S_{inf})e^{-Br}$$

In this case, the ratio $S_0 / S_{\inf}$ is set to 10 to create a clustered scenario, whilst exponent B describes the number of farms in a radius. When $B = 0$ farms are randomly distributed, as B increases the local density increases. To create the clustered farm demography, B is set to 0.5 and the equation above is used to determine the density of farms around each farm in the landscape at all distances *r* within the domain. Sensitivity to these assumptions is examined in detail in [21].

2.5 EPIDEMIOLOGICAL BEHAVIOUR

The main aim of this study is to investigate the feasibility of using a metapopulation model to predict epidemic dynamics in a highly heterogeneous landscape; we compare key epidemiological parameters (such as epidemic size and spatial spread) between the three different models. The IFB model and the metapopulation models are used to simulate outbreaks in Cumbria, Devon and Aberdeenshire. Cumbria and Devon played a significant role in the 2001 FMD epidemic whilst Aberdeenshire was chosen owing to the large number of farms and high farm density in Scotland.

Final epidemic size is one of the key epidemiological parameters. To test how well the metapopulation models perform in predicting the epidemic size compared with the IFB model, we initiate 10,000 epidemics in each of the three studied counties, with one random index case, and record the final epidemic size for each simulated epidemic.

To examine how well the metapopulation models perform compared with the IFB model in predicting spatial spread and initiate epidemics in a single farm in the North of Devon and North-East of Aberdeenshire. We record the average spatial extent in Devon and Aberdeenshire of all models. As we are most interested in the characteristics of large epidemics we select 200 large epidemics (>1000 infected farms for Devon; >500 infected

farms for Aberdeenshire) for each model and grid cell size. This eliminates epidemics that die out owing to early stochastic extinction. The outbreaks in the metapopulation models are initiated with a single farm in the grid cell where the index case of the IFB is located and the prevalence on a within an individual grid cell is recorded and averaged over all runs. To compare the spatial extent between the IFB and metapopulation models, we aggregated the simulation outputs of the IFB according to the corresponding grid cell in the metapopulation model. We compare the mean proportion of infected farms in grid cell k when the IFB is applied with both versions of the metapopulation model with grid cell sizes of 200 metres to 10 km.

3. RESULTS

3.1 OUTBREAK SIZE

Figure 1 shows the distribution of final epidemic sizes for Devon for the IFB model and the two metapopulation models. The IFB model predicts the smallest epidemics on average with the highest probability of early stochastic fade-out for all three models and all three counties studied. Outbreaks initiated in Devon (figure 1, figure S1) and Aberdeenshire (figure S1-S2) display strikingly different risks of stochastic extinction and outbreak sizes, compared to those of Cumbria (figure S1 figure S3).

Both versions of the metapopulation model overestimate the probability of take-off and the epidemic size in each of the counties. However, the heterogeneous model, for grid cell sizes ≤ 5 km, correctly predicts that Cumbria has the greatest risk of epidemic take off, followed by Devon and then Aberdeenshire (figure S1). The homogeneous model only proves accurate in this regard for grid cell sizes ≤ 1 km. We also observe the smallest variation in epidemic sizes in Cumbria in the IFB model, with wider distributions found in the other two counties.

3.2 SPATIAL DISTRIBUTION

When seeding an epidemic in the north of Devon, the results for all three models indicate that the outbreak remains most concentrated in the northern area where the epidemic is seeded (figure 2). The heterogeneous metapopulation model (figure 2E-G) outperforms the homogeneous version (figure 2B-D) in that it predicts similar spatial spread to the IFB (figure 2A). However, the spatial spread predicted by both metapopulation models, and all tested grid cell resolutions, shows high agreement with the IFB for Devon, Aberdeenshire and Cumbria (figure 2H-J, figure S4-5H-J).

Our results indicate that the accuracy of metapopulation models in terms of estimating mean grid cell prevalence is correlated to the number of farms per grid cell (figure S6) – the more farms in a grid cells the more likely that the mean prevalence will be overestimated. The homogeneous metapopulation model (4 km, figure S6A) over predicts the prevalence by at least 10% compared with the IFB in 52% of the grid cells. When the heterogeneous model is used, only 28% of grid cells overestimate (>10%) the grid cell prevalence when compared with the IFB (4 km, figure S6B). With the lowest grid cell solution tested (10km), 43% (homogeneous, figure S6C) and 30% (heterogeneous, figure S6D) of the grid cells overestimate the mean grid cell prevalence.

Whilst the metapopulation models slightly over-predict the mean grid cell prevalence, they may still prove to be useful tools for policy makers, provided that the models do not underpredict the impact of spatial spread. There is an inherent risk associated with utilizing models that tend to underpredict epidemic extent – any recommended intervention strategy may be insufficient to control an outbreak, leading to an increased likelihood of a large scale epidemic occurring. Should a model over-predict the size of an outbreak, any intervention strategy may be too draconian and result in a slight increase in the overall number of farms affected, but such a control policy should eradicate the epidemic.

The risk of underprediction is more apparent for the simulations in Devon (figure 2) and Cumbria (figure S5) than Aberdeenshire (figure S4). In Devon and Cumbria, underprediction of the proportions of infected farms in a grid cells are more likely to occur in grid cells of 10 km (figure 2J) – this is seen in both the homogeneous and the heterogeneous metapopulation model. In 11 (homogeneous) and 10 (heterogeneous) of the 87 grid cells of 10 km by 10 km, the grid cell prevalence is underestimated by at least 10%. Grid cells with a low number of

farms but with a high variation in farm size and/or composition (and therefore a large variation in farm susceptibility within the grid cell) are most likely to underestimate the mean grid cell prevalence when compared with the IFB model. However, the metapopulation models with a grid cell size of 4 km perform substantially better. The prevalence per grid cell of a small proportion of individual grid cells (2% for the homogeneous model, 3% for the heterogeneous model) underestimates the overall grid cell prevalence by at least 10%.

The absolute difference of the prevalence between the metapopulation models and IFB are linked to the relative proportion of grid cells that do not underestimate the grid cell prevalence (error of 10%, figure 2H-J). For Devon only, the homogeneous model with a grid cell size of 10 km appears to perform better than the 4 km in predicting the epidemic size (figure 1). However, the proportion of infected farms is overpredicted in some grid cells and underpredicted in others at this larger scale. This grid cell-level inaccuracy is not highlighted by simply focusing on the final epidemic sizes at the county scale, therefore, considering only the predicted epidemic size appears to be insufficient to examine the accuracy of a model (figure S7). This indicates that the optimal grid cell resolution is 4km for both the homogeneous and heterogeneous models (2H-J, S4H-J, S5H-J, S7).

One of the benefits of using a metapopulation model over an IFB model is the reduced computational time. The time needed to run simulations rapidly declines as the size of the grid cells increases (figure S8); as an example, simulations at grid cell sizes of 5km only take 6.3% of the time needed to simulate using 200m grid cells.

3.3 SPATIAL CLUSTERING OF FARMS

We investigate how well the metapopulation models perform in a situation where there is no spatial clustering of farms and compared this with a scenario where farms are very clustered. We run 2500 simulations for both clustering scenario and grid cell resolutions. Clustering has a substantial effect on the final epidemic size – the more clustered the farms, the higher the average final epidemic size of the IFB. When farms are randomly distributed, average epidemic size is 625 farms (figure 3A) compared with 1932 farms when the farms are very clustered (figure 3B). The homogeneous and heterogeneous models both perform better

in the random farm location scenario than in the clustered situation. Epidemic sizes are best replicated for grid cell sizes ≤ 4 km when the farms are randomly allocated. When the grid cell resolution decreases the epidemic size predictions are overestimated, as seen in the epidemic size predictions for the three studied counties.

In the clustered situation we notice the opposite effects on the epidemic size predictions of the metapopulation models. When the grid cells resolution decreases the metapopulation models underestimate the epidemic size. Therefore, in extremely clustered situations, high-resolution grids would be needed to avoid the risk of underestimating the impact of an epidemic.

4. DISCUSSION

This study investigates the robustness of metapopulation models for predicting epidemiological behaviour compared with an IFB model approach. Within a patch in a metapopulation model, all farms are considered uniform and therefore the effect of heterogeneity of farm size, composition and spatial location are more difficult to capture. Previous studies have identified the importance of heterogeneity of farm size on epidemiological behaviour [1,21-22]. However, this is the first study that identifies how well a metapopulation performs in predicting the spatial spread and final epidemic size when compared with an IFB model. Moreover, this study identifies the level of spatial aggregation that can be utilised in a metapopulation model in order to maintain accurate epidemiological predictions.

The metapopulation models generally overestimate epidemic size when compared with the IFB whilst a heterogeneous model that includes farm sizes is found to perform better than a homogeneous model. The heterogeneous metapopulation model predicts the spatial spread, risk for epidemic take-off and the relative prevalence between counties correctly for grid cells with a lower resolution than the homogeneous model. Overestimation of epidemic size may not always be a problem for models that are used to inform policy – policy makers are often most concerned in the worst-case scenario of a potential outbreak and therefore it is important that a simulation model captures the spatial spread of an epidemic.

Whilst the homogenous model performs well at predicting final epidemic size for grid cells up to 10km, this version of the model underestimates the grid cell prevalence in a substantial proportion of grid cells. Even if the heterogeneous version is used the proportion of grid cells where the grid cell prevalence is underestimated does not improve. The 4km grid cells perform substantially better than the 10km grid cells. Therefore, in cases where a metapopulation model is preferred over an IFB model owing to data limitations, using the heterogeneous metapopulation model with grid cells of 4 km provides the most optimal balance between privacy and detailed model outputs. However, in very clustered landscapes the metapopulation models are more likely to underestimate the final epidemic size and spread. This all assumes parameters are known and fixed. In practice parameters may reflect and compensate for model deficiencies; this is true for both IFB and metapopulation models. Therefore when using metapopulation models in a practical setting and fitting the model to data many of these differences may disappear. However, if it is not possible to determine specific farm locations, information regarding local clustering of farm populations is crucial to enable accurate local-scale predictions of future epidemic behaviour.

In cases where IFB data are unavailable it may be possible to estimate farm locations using other geographic information such as land cover data that are available in the public domain [26]. Some countries have a dataset available that includes an estimated number of farms in a province or county – these data could be used in combination with landscape data to produce a synthetic dataset. In situations such as these, the use of synthetic location data and aggregated information on farm size and species composition, coupled with a metapopulation model framework, may allow for an investigation into the effectiveness of control strategies for livestock disease outbreaks at a coarse spatial scale.

The model framework detailed in this paper is not only able to capture FMD, but may also be applicable to other infectious diseases such as classical swine fever and avian influenza. The distant dependent transmission kernel could be fitted to outbreak data in a similar way as has been performed for FMD [24]. Live animal movements could also play a role in disease transmission, especially during the silent spread phase prior to detection of disease and may cause the disease to be geographical widespread. These movements can be easily included in both the IFB model and the metapopulation model [8]. This paper shows that stratifying

farm sizes in a metapopulation framework could offer benefits over a standard metapopulation framework. We expect that there is potential for applying this heterogeneous model to human disease spread and in other animal diseases. Furthermore, in situations where the number of farms is very large (such as in the USA and Australia), aggregating the data to feed into a metapopulation model may be more practical and will reduce the computational power considerably.

The results of this paper suggest that there may be a preference for an IFB model over a metapopulation model when it is important to accurately quantify characteristics such as final epidemic size. However, in many cases an understanding of the relative behaviour of outbreaks in different geographic regions and an accurate prediction of the spatial spread of a potential outbreak is of equal importance. Moreover, in many situations an IFB model is simply not feasible, as data to inform such a model simply do not exist. In such circumstances a heterogeneous metapopulation model could be utilised to predict epidemic behaviour. Understanding the limitations of using a metapopulation model on predicting key epidemiological quantities is vital to make these types of models useful for policy makers in the event that an IFB model cannot be used. Sensitivity analysis of disease parameters and the shape and scale of the transmission kernel could improve the robustness of the qualitative outcome of the model, such as which control strategies should be implemented and which geographic regions are most likely to result in a high risk of large scale epidemics occurring.

ACKNOWLEDGEMENTS

This work was made possible by funding from the Research and Policy for Infectious Disease Dynamics (RAPIDD) Program, Science and Technology Directorate, US Department of Homeland Security and Fogarty International Center, National Institutes of Health and the Foreign Animal Disease Modeling Program, Science and Technology Directorate, US Department of Homeland Security (Grant ST-108-000017) and BBSRC (grant no. BB/K010972/3). MJK was funded through the ERA-NET anihwa scheme with funding provided by the Department for Environment, Food & Rural Affairs (Defra). Data was kindly provided by Defra. We are thankful to Colleen Webb and Uno Wennergren for useful discussion on this manuscript.

REFERENCES

1. Keeling, MJ, Woolhouse, ME, Shaw, DJ, Matthews, L, Chase-Topping, M, Haydon, DT, Cornell, SJ, Kappey, J, Wilesmith, J, Grenfell, BT . 2001 Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. Science 294: 813-817. (doi: 10.1126/science.1065973)

2. Ferguson, NM, Donnelly, CA, Anderson, RM. 2001 The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. Science 292: 1155-1160. (doi: 10.1126/science.1061020)

3. Ferguson, NM, Cummings, DAT, Fraser, C, Cajka, JC, Cooley, PC, Burke, DS. 2006 Strategies for mitigating an influenza pandemic. Nature 442: 448-452. (doi:10.1038/nature04795)

4. Germann, TC, Kadau, K, Longin, IM, Macken, CA. 2006 Mitigation strategies for pandemic influenza in the United States. Proc Natl Acad Sci U S A 103: 5935-5940. (doi: 10.1073/pnas.0601266103)

5. Brooks-Pollock, E, Roberts, GO, Keeling, MJ. 2014 A dynamic model of bovine tuberculosis spread and control in Great Britain. Nature 511: 228-231. (doi:10.1038/nature13529)

6. Webb, CT, Ferrari, M, Lindstrom, T, Carpenter, T, Durr, S, Garner, G, Jewell, C, Patyk, KA, Stevenson, M, Ward, MP, Werkman, M, Tildesley, MJ “in prep” Ensemble Modeling and structured decision-making to support emergency disease management. In prep.

7. Keeling, MJ, Danon, L, Vernon, MC, House, TA. 2010 Individual identity and movement networks for disease metapopulations. Proc Natl Acad Sci U S A 107: 8866-8870. (doi: 10.1073/pnas.1000416107)

8. Buhnerkempe, MG, Tildesley, MJ, Lindstrom, T, Grear, DA, Portacci, K, Miller, RS, Lombard, JE, Werkman, M, Keeling, MJ, Wennergren, U, Webb, CT. 2014 The Impact of Movements and Animal Density on Continental Scale Cattle Disease Outbreaks in the United States. PLoS ONE 9: e91724. (doi: 10.1371/journal.pone.0091724)

9. Tildesley, MJ, Bessell, PR, Keeling, MJ, Woolhouse, ME. 2009 The role of pre-emptive culling in the control of foot-and-mouth disease. Proc Biol Sci 276: 3239-3248. (doi: 10.1098/rspb.2009.0427)

10. Tildesley, MJ, Savill, NJ, Shaw, DJ, Deardon, R, Brooks, SP. 2006 Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. Nature 440: 83-86. (doi:10.1038/nature04324)

11. Buhnerkempe, MG, Grear, DA, Portacci, K, Miller, RS, Lombard, JE, . 2013 A national-scale picture of U.S. cattle movements obtained from Interstate Certificate of Veterinary Inspection data. Prev Vet Med 112: 318-329. (doi: 10.1016/j.prevetmed.2013.08.002)

12. Garner, M and Beckett S (2005) Modelling the spread of foot-and-mouth disease in Australia. Aust Vet J 83: 758–766. doi: 10.1111/j.1751-0813.2005.tb11589.x

13. Levin, SA, Grenfell, B, Hastings, A, Perelson, AS. 1997 Mathematical and Computational Challenges in Population Biology and Ecosystems Science. Science 275: 334-343. (doi: 10.1126/science.275.5298.334)

14. Hanski, I. 1999 Metapopulation ecology. New York: Oxford University Press., 328 pp. (doi: 10.1098/rstb.1995.0070)

15. Bolker, B, Grenfell, B. 1995 Space, Persistence and Dynamics of Measles Epidemics. Philos Trans R Soc Lond B Biol Sci 348: 309-320.

16. Hanski, I. 1994 A practical model of metapopulation dynamics. J Anim Ecol 63: 151-162. (doi: 10.2307/5591)

17. Keeling, MJ, Gilligan, CA. 2000 Bubonic plague: a metapopulation model of a zoonosis. Proc. R. Soc. Lond. B 267: 2219-2230. (doi: 10.1098/rspb.2000.1272)

18. Gaff, HD, Gross, LJ. 2007 Modeling Tick-Borne Disease: A Metapopulation Model. Bull Math Biol 69: 265:288. (doi: 10.1007/s11538-006-9125-5)

19. Green, DM, Kiss, IZ, Kao, RR. 2006 Modelling the initial spread of foot-and-mouth disease through animal movements. Proc. R. Soc. B 273: 2729–2735. (doi: 10.1098/rspb.2006.3648)

20. Gibbens, JC, Sharpe, CE, Wilesmith, JW, Mansley, LM, Michalopoulou, E, Ryan, JBM, Hudson, M . 2001 Descriptive epidemiology of the 2001 foot and mouth disease epidemic in Great Britain: The first five months. Vet Rec 149: 729–743. (doi:10.1136/vr.149.24.729)

21. Tildesley, MJ, House, TA, Bruhn, MC, Curry, RJ, O'Neil, M, Allpress, JLE, Smith, G, Keeling, MJ. 2010 Impact of spatial clustering on disease transmission and optimal control. Proc Natl Acad Sci U S A 107: 1041-1046. (doi: 10.1073/pnas.0909047107)

22. Rock, K, Brand, S, Moir, J, Keeling, MJ. 2014 Dynamics of infectious diseases. Rep Prog Phys 77: 026602. (doi: 10.1088/0034-4885/77/2/026602)

23. Tildesley, MJ, Deardon R, Savill, NJ, Bessell, PR, Brooks, SP, Woolhouse, ME, Grenfell, BT, Keeling, MJ (2008). Accuracy of models for the 2001 foot-and-mouth epidemic. Proc. R. Soc. B 275: 1459–1468. (DOI: 10.1098/rspb.2008.0006)

24. Rorres, C, Pelletiera, STK, Keeling, MJ, Smith, G. 2010 Estimating the Kernel Parameters of Premises-Based Stochastic Models of Farmed Animal Infectious Disease Epidemics using Limited, Incomplete, or Ongoing Data. Theor Popul Biol 78: 46-53. (doi: 10.1016/j.tpb.2010.04.003)

25. Ryan, E, Gloster, J, Reid, SM, Li, Y, Ferris, NP, Waters, R, Juleff, N, Charleston, B, Bankowski, B, Gubbins, S et al. 2008 Clinical and laboratory investigations of the outbreaks of foot-and-mouth disease in southern England in 2007. Vet Rec 163: 139-147. (doi:10.1136/vr.163.5.139)

26. Tildesley, MJ, Ryan, SJ. 2012 Disease Prevention versus Data Privacy: Using Landcover Maps to Inform Spatial Epidemic Models. PLoS Comput Biol 8: e1002723. (doi: 10.1371/journal.pcbi.1002723)

FIGURE CAPTIONS

Figure 1: The distribution of the epidemic extent for grid cell sizes of 200 metres, 1, 4 and 10 km for homogeneous (A) and heterogeneous (B) metapopulation model and the individual farm-based model for Devon.

Figure 2: Epidemic extent of outbreaks in the north of Devon. The same index case was used and 200 iterations were selected that had more than 1000 infected farms. The colour of the dot/ grid cell represents how often a farm becomes infected. The leftmost figure in the middle shows the spatial spread when the individual farm-based model is used (A). The top graphs show the results for the heterogeneous model with a 1km (B), 4km (C) and 10km (D) resolution and the graphs in the middle show the results of the homogeneous model, for 1 km (E), 4km (F) and 10 km (G) resolution. H-J show the correlation between the mean proportion of infected farms in grid cell k in the metapopulation model compared with the individual farm-based model for grid cells of 1 km (H), 4 km (I) and 10 km (J). Blue dots show the correlation between the individual farm-based and heterogeneous model and the red dots show the results for the homogeneous model. The white (H) and black (I-J) dashed line represents $x = y$.

Figure 3: The accuracy of metapopulation models with different spatial clustering of farms. Situation A: farms are randomly distributed (situation A). Situation B: farms are extremely clustered, $S_0 / S_{\inf} = 10$ and $B = 0.5$. The left panels show the locations of farms ($n = 2500$) in a 50 by 50 km^2. The right panels show the final average size when the individual farm-based, heterogeneous and homogenous models are used.

Figure S1: Likelihood of epidemics taking off (more than 10 infected farms) in three different counties: Cumbria, Devon and Aberdeenshire for the homogeneous model (HO) and heterogeneous model (HE). The horizontal lines represent the likelihood of take-off when the individual farm-based model is used for all three counties.

Figure S2: The distribution of the epidemic extent for grid cell sizes of 200 metres, 1, 4 and 10 km for homogeneous (A) and heterogeneous (B) metapopulation model and the individual farm-based model for Aberdeenshire.

Figure S3: The distribution of the epidemic extent for grid cell sizes of 200 metres, 1, 4 and 10 km for homogeneous (A) and heterogeneous (B) metapopulation model and the individual farm-based model for Cumbria.

Figure S4: Epidemic extent of outbreaks in the centre of Aberdeenshire. The same index case was used and 200 iterations were selected with more than 500 infected farms. The colour of the dot/ grid cell represents how often a farm becomes infected. The leftmost figure in the middle shows the spatial spread when the individual farm-based model is used (A). The top graphs show the results for the heterogeneous model (B-D) and the graphs in the middle show the results of the homogeneous model, for 1 km (E), 4km (F) and 10 km (G) grid cells. H-J show the correlation between the mean proportion infected farms in grid cell k in the metapopulation model compared with individual farm-based model for grid cells of 1 km (H), 4 km (I) and 10 km (J). Blue dots show the correlation between the individual farm-based and heterogeneous model and the red dots show the results for the homogeneous model. The white (H) and black (I-J) dashed line represent $x = y$.

Figure S5: Epidemic extent of outbreaks in the centre of Cumbria. The same index case was used and 200 iterations were selected with more than 500 infected farms. The colour of the dot/ grid cell represents how often a farm becomes infected. The leftmost figure in the middle shows the spatial spread when the individual farm-based model is used (A). The top graphs show the results for the heterogeneous model (B-D) and the graphs in the middle show the results of the homogeneous model, for 1 km (E), 4km (F) and 10 km (G) grid cells. H-J show the correlation between the mean proportion infected farms in grid cell k in the metapopulation model compared with individual farm-based model for grid cells of 1 km (H), 4 km (I) and 10 km (J). Blue dots show the correlation between the individual farm-based and

heterogeneous model and the red dots show the results for the homogeneous model. The white (H) and black (I-J) dashed line represent $x = y$.

Figure S6: This graph shows the difference in mean grid cell prevalence between the individual farm-based model and the metapopulation model, explained by the number of farms in the grid cell and the maximum range of susceptibility in the grid cell. Figures A-B show the result for grid cell of 4km figures C-D show the results for 10 km. Figure A&C shows the results for the homogeneous metapopulation model and B&D show the heterogeneous version.

Figure S7: This graph shows the increase in residuals (squared difference between the metapopulation models and individual farm-based based model for grid cell prevalence, left) for the homogeneous model (red) and heterogeneous model (blue) and the proportion of grid cells that are not underestimated (right).

Figure S8: Mean computational time of one simulation (averaged over 1000 simulations) for the homogeneous and heterogeneous metapopulation model of each grid cell size.

Table S1. Values for the key parameters used in this model.

FIGURE 1

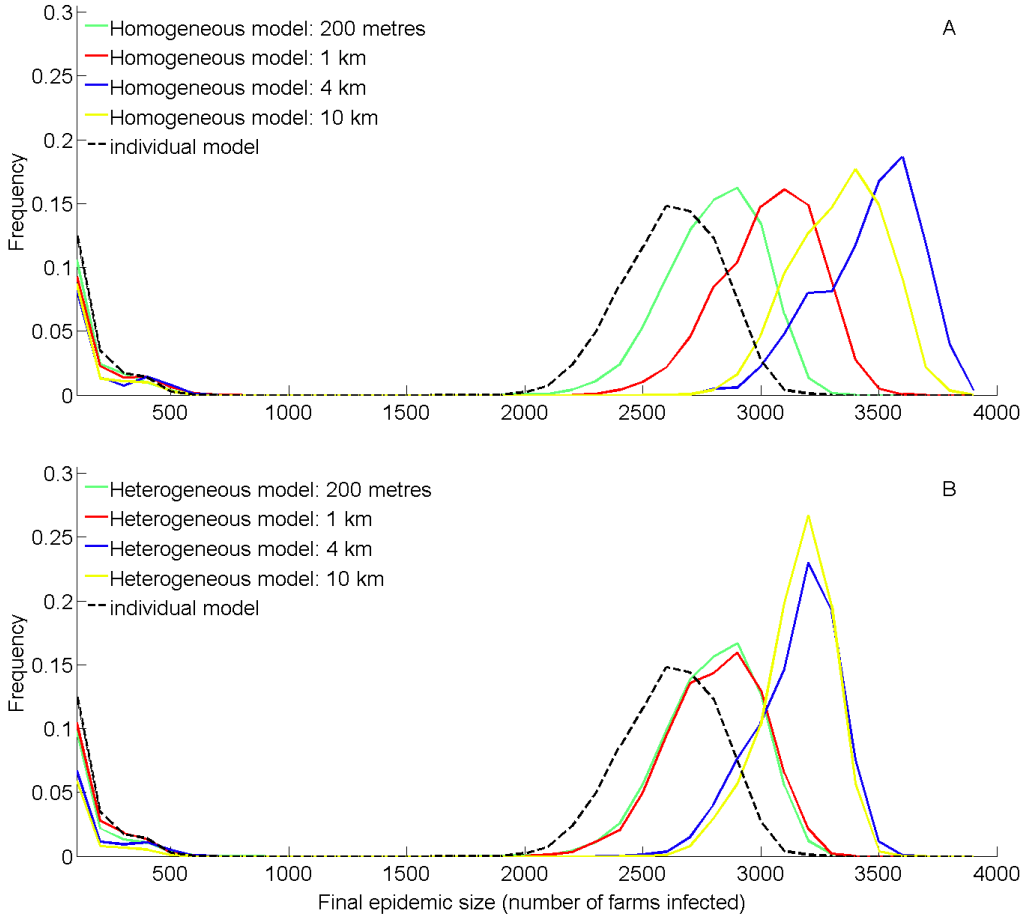

Figure 1: The distribution of the epidemic extent for grid cells sizes of 200 metres, 1, 4 and 10 km for homogeneous (A) and heterogeneous (B) metapopulation model and the individual farm-based model for Devon.

FIGURE 2

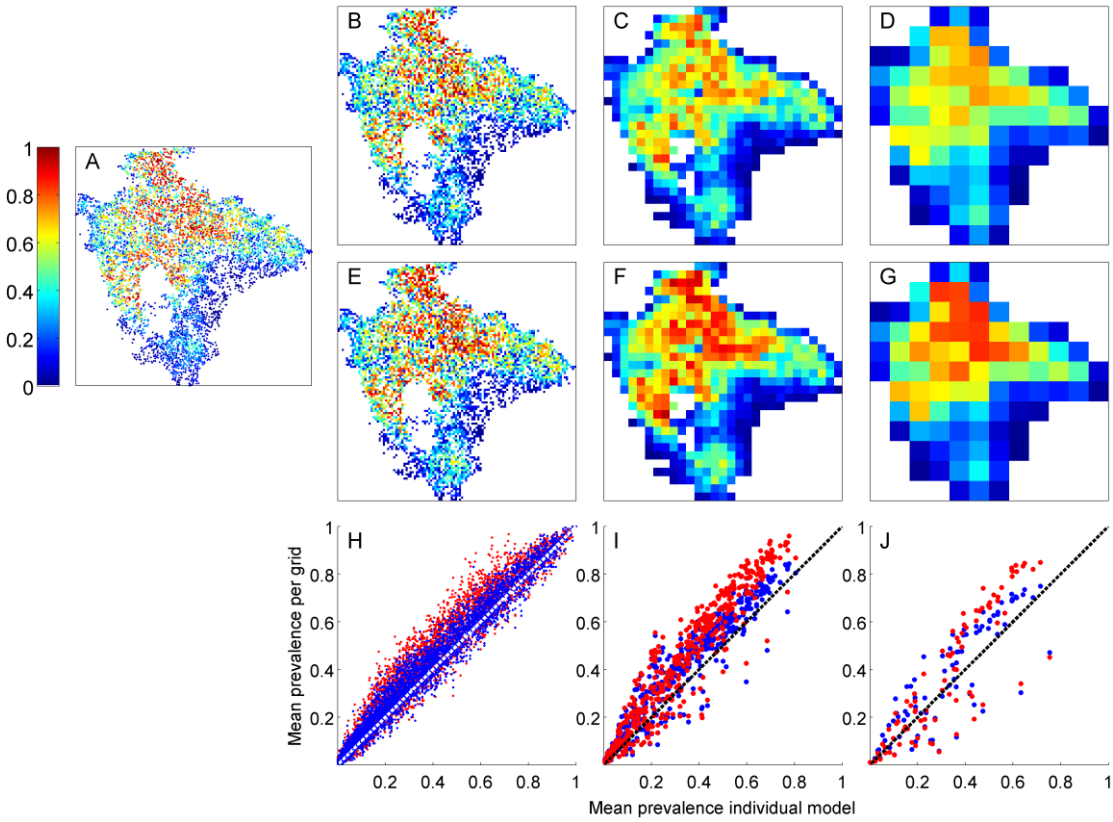

Figure 2: Epidemic extent of outbreaks in the north of Devon. The same index case was used and 200 iterations were selected that had more than 1000 infected farms. The colour of the dot/ grid cell represents how often a farm becomes infected. The leftmost figure in the middle shows the spatial spread when the individual farm-based model is used (A). The top graphs show the results for the heterogeneous model with a 1km (B), 4km (C) and 10km (D) resolution and the graphs in the middle show the results of the homogeneous model, for 1 km (E), 4km (F) and 10 km (G) resolution. H-J show the correlation between the mean proportion of infected farms in grid cell k in the metapopulation model compared with the individual farm-based model for grid cells of 1 km (H), 4 km (I) and 10 km (J). Blue dots show the correlation between the individual farm-based and heterogeneous model and the red dots show the results for the homogeneous model. The white (H) and black (I-J) dashed line represents $x = y$.

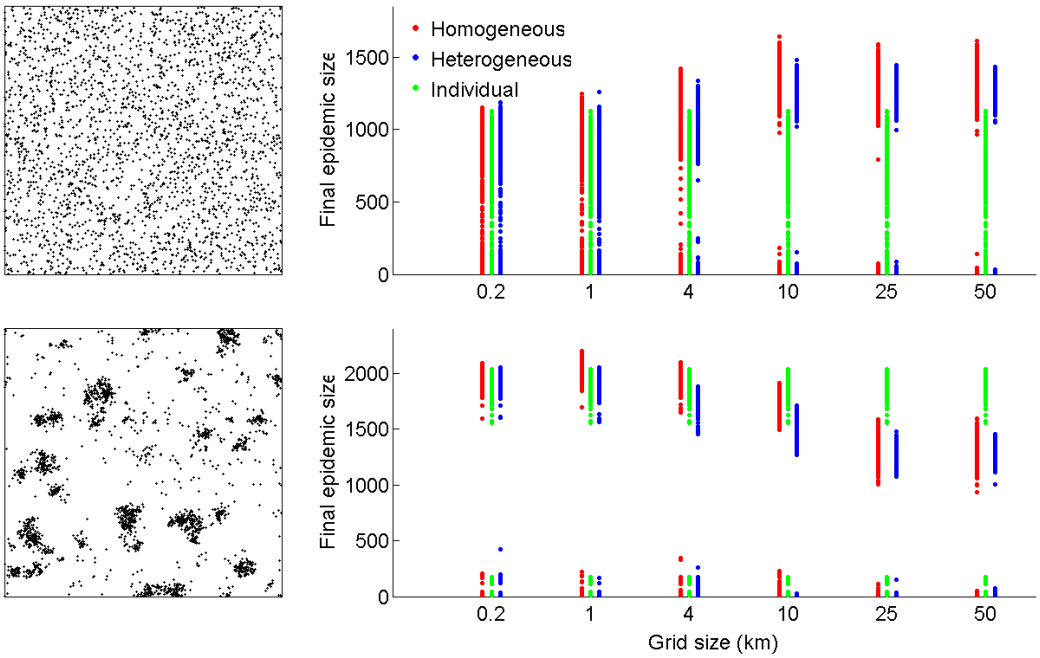

Figure 3: The accuracy of metapopulation models with different spatial clustering of farms. Situation A: farms are randomly distributed (situation A). Situation B: farms are extremely clustered, $S_0 / S_{\inf} = 10$ and $B = 0.5$. The left panels show the locations of farms ($n = 2500$) in a 50 by 50 km^2. The right panels show the final average size when the individual farm, heterogeneous and homogenous models are used.